基于 K 近邻的众包数据分类算法 *

李佳烨 [1]，余 浩 [2†]

(1. 广西师范大学 广西多源信息挖掘与安全重点实验室，广西 桂林 541004; 2. 中南大学 信息科学与工程学院，长沙 410083)

摘 要：针对众包数据处理中的质量控制问题，提出了一种加权 K 近邻投票分类方法。该方法不单单只是考虑了某个样例的标记来返回一个答案，而是通过综合考虑样例的近邻来得到更加准确的答案。同时对样例的近邻加以适当的权重来进一步提高算法的性能，并保持了传统多数投票分类的简单性。K 近邻投票分类算法可以有效地解决缺乏标记的情况，通过对近邻加以权重可以解决不平衡标记造成的影响，从而使算法的泛化性更强。通过各种场景下的实验，结果表明加权 K 近邻投票分类方法取得了很好的效果。

关键词：众包数据；质量控制；K 近邻投票；多数投票

中图分类号：TP181　　doi: 10.19734/j.issn.1001-3695.2018.09.0736

Crowdsourcing data classification algorithm via K-nearest neighbor

Li Jiaye[1], Yu Hao[2†]

(1. *Guangxi Key Laboratory of Multi-source Information Mining & Security*, *Guangxi Normal University*, *Guilin Guangxi* 541004, *China*; 2. *Central South University*, *School of Information Science & Engineering*, *Changsha* 410083, *China*)

Abstract: Aiming at the quality control problem in crowdsourcing data processing, this paper proposed a weighted K-nearest neighbor voting method. This method not only considers the mark of a certain sample to return an answer, but rather obtains a more accurate answer by considering the neighbors of the sample comprehensively. At the same time, it applies appropriate weights to the neighbors of the sample to further improve the performance of the algorithm and maintain the simplicity of the traditional majority vote. The K-nearest neighbor vote can effectively solve the problem of lack of markup. By weighting the neighbors, it can solve the influence of the unbalanced mark. And the generalization of the algorithm is stronger. Through experiments in various situations, the results show that the proposed weighted K-nearest neighbor voting method has achieved good results.

Key words: crowdsourcing; quality control; K-nearest neighbor voting; majority voting

0 引言

随着人工智能时代的到来，数据的重要性已经不言而喻[1]，并且影响世界的方方面面。阿里巴巴的城市大脑项目将 AI 技术应用在系统收集到的城市大数据上，可以 20 min 锁定嫌犯；滴滴出行通过对城市大数据分析来为每辆车规划最优行车路线，缓解交通拥堵问题；沃尔玛通过超市顾客的选购记录数据分析，为商家精准投放广告。这些数据收集的任务大部分都可以使用机器自动完成，但是在图像类别标注[2]，商品优劣等一些任务中，机器常常无法准确地对这些数据进行处理。近年来，有学者在研究中发现数据的有效性与数量级对实验效果的影响甚至高于对算法本身进行优化[3]。因此，如何在研究的相关领域找到高质量与高数量级兼具的数据集成为了众多科研工作者迫切需要解决的问题。2009 年普林斯顿大学李飞飞团队推出了一个 ImageNet 的图像数据集[4]。目前是世界上最大的图像识别数据库，经过近十年的发展，ImageNet 数据集已经对计算机视觉以及整个机器学习领域带来了深远的影响。在 2010—2017 年期间，研究人员已经将物体分类的准确率提高到了 97.3%，超过了人类分辨水平[5]。随着时间推移，这个数字仍然在一步步向 100%靠近。可以说 ImageNet 数据集打开了计算机视觉，模式识别甚至整个人工智能领域的研究新篇章，同时也充分证明了巨量级的真实数据集对科学研究的重要性。在 ImageNet 数据集研发之初，如何对 1 600 万张图像数据进行标注成为了关键难点。在当时看来这个几乎是个不可能完成的任务，团队创始人李飞飞偶然接触到 amazon 的众包平台 Amazon Mechanical Turk，通过众包平台将大量的图片标注任务分配给世界上任何一个与之感兴趣的人，最终仍然经过了两年多的时间才完成了这个标注任务。ImageNet 数据集的成功也同时证明了利用众包方法处理一些问题是非常必要与高效的。然而众包方法便利地带来大量数据的同时也带来了一些问题。通常很多标记人员并不是相关的专业人士，水平参差不齐，每个人给出的数据标注并不是完全正确的，有些把对象标注了错误的标签，有些因为不确定而没有给出标签。这些缺陷导致最终得到的众包数据通常含有一定的噪声值与缺失值。这些问题在数据挖掘中非常常见[6]，因此，如何针对性地处理这些问题，成为能否高效利用众包数据的关键。

在之前 Zhang 等人[7]提出了 Efficient kNN Algorithm

收稿日期：2018-09-26；**修回日期：**2018-11-16　**基金项目：**国家重点研发计划资助项目(2016YFB1000905)；国家自然科学基金资助项目(61170131，61263035，61573270，90718020)；国家“973”计划资助项目（2013CB329404）；中国博士后科学基金资助项目（2015M570837）；广西自然科学基金资助项目（2015GXNSFCB139011，2015GXNSFAA139306）

作者简介：李佳烨（1993-），男，山西晋城人，硕士研究生，主要研究方向为数据挖掘、机器学习；余浩（1994-），男（通信作者），江西上饶人，硕士，主要研究方向为数据挖掘、机器学习（yhgxnu@gmail.com）.

Based on Graph Sparse Reconstruction 来对传统的 KNN 算法做了改进。它通过 l_1－范数来动态的为不同的样本生成不同的 K 值，以此来使算法达到很好的性能。但是这种算法只能适用于一般的数据集，并不能用于众包数据，而且它对数据标记的缺失很敏感。此外，郝建柏等人[8]也提出了一种模糊近邻标签传递的半监督分类算法。该方法虽然能对没有标签的数据进行分类，但是它主要用于半监督学习，且只能进行单标签的学习，对于众包数据的多标签无法进行分类，因此也不适用于众包数据。

本文针对众包数据中出现的这些问题，提出了加权 K 近邻投票分类算法。具体地，应用 K 近邻投票分类方法解决数据标签缺失的情况，同时通过赋予近邻适当的权重来更加合理地对算法进行改进。最终同时结合以上两种改进方法，提出了加权 K 近邻投票分类算法。

1　相关理论

1.1　KNN

KNN 算法指的是 K 最近(K-nearest neighbour)算法[9]，是数据挖掘众多算法中最为经典的算法之一。kNN 算法的主要思想是数据集中某一点的类别可以由其周围的 k 个已被正确分类的点来决定，也就意味着数据中的任意一点与其靠近的点具有相同的类别。具体地，首先设定 k 值，然后通过计算该点与其他所有点的距离，之后保留距离小的 k 个邻居样本点，最后通过计算这 k 个点中出现概率最大的类别，即该点的预测类别。距离的度量方式主要包括两种，一种是曼哈顿距离，另一种是欧氏距离。由于 Knn 方法判定类别是根据其周围有限的 K 个点来决定的，而与整体类别的关系不大，所以该方法尤其适合处理样本类别重叠较多数据集的多分类问题。很多学者在 knn 算法基础上进行了很多改进研究。为了解决 KNN 算法容易受 k 值的影响的问题，Zhang 等人[10]提出了一种为不同测试数据设定不同 k 值的 CM-Knn 算法。由于 KNN 依赖样本类别平衡，Zhang[11]提出了一种结合确定性因素的 KNN-CF 算法。KNN 算法因为其算法思路简单，实现难度低，分类效果好，同时适合处理多分类问题而在数据挖掘领域被广泛应用。

1.2　随机森林算法

随机森林算法是一种基于决策树改进的机器学习算法[12]。同时决策树[13]是监督学习领域中一个非常经典的分类算法。与传统的逻辑回归分类算法不同，决策树算法是一个树型模型（非线性模型）。传统的逻辑回归算法是通过将所有特征进行加权处理得到一个值，然后某一个设定的阈值将样本线性划分。而决策树通过生成的树型结构，将所有特征分开处理，从而做到对样本更加精确的非线性分割。决策树模型中根节点代表最主要的特征，每个子节点代表特定的区分特征，最下层的叶节点则代表样本的最终类别。其主旨思想是将特征差别大的样本尽可能地分开。随机森林算法通过将集成学习方法应用在决策树，将多个决策树的投票结果组成一个强分器，以此来提高了准确率。随机森林算法有如下几种特点：随机抽样训练集，使得每棵树接受不相同的训练样本，从而提高了模型的泛化性；进行有放回的抽样方式，使得每棵树的抽取训练样本的方式既有随机，同时也有相关性，从而能够获取到内部生成误差的一种无偏估计。由于随机森林算法具有高准确率、适合处理高维数据、对缺失值不敏感等特点被广泛应用在推荐系统、预测模型等方面。

2　算法描述

2.1　MV 方法

如表 1 所示，可以把它看做一个众包数据，其中，甲、乙、丙代表标记员，X1,X2,…,Xn 表示第一个样例到第 n 个样例。表中的 L 数据代表标记员对样例进行的标记。在实际工作中，众包数据难免会有部分标记缺失。同时，存在标记员工作懈怠，或专业水平有限导致标记结果出现错误。这些都是众包数据中存在的问题。

表 1　众包数据

Table 1　Crowdsourcing data

	X1	X2	……	Xn
甲	L11	L12	……	L1n
乙	L21	L22	……	L2n
丙	L31	L32	……	L3n

为了从众包数据中得到正确的标记，最经典的 MV(majority voting)方法[14]根据少数服从多数的原则来实现。具体如式（1）所示。

$$v(x) = \arg\max_{c\in\Omega}\ v(c\,|\,x) \tag{1}$$

其中：$v(c\,|\,x) = \frac{1}{|S_x|}\sum_{l\in S_x}\mathbf{1}(l = \mathrm{c})$；$|S_x|$ 是指标记员的个数；S_x 是一个样例对应的所有标记（是一个集合，里面的标记个数等于标记员的个数）；$\mathbf{1}(\cdot)$ 表示：若 ( ) 中的内容为真，则返回 1，否则返回 0；l 是标记；c 是真正的类别；$\Omega=\{1,\ldots\ldots,c\}$ 是标签集合。很明显，若 $v(c\,|\,x)>0.5$，则表示根据 MV 方法得到的这个样例的标记是正确的。

虽然此方法能达到一定的效果，但是它默认的每个标记员的知识水平或能力是一样的[15]。比如表 1 中，若甲为本科文凭，且能力很强；乙为高中文凭，能力较强；丙为初中文凭，能力一般。此时若甲标记正确，乙丙两人标记错误，根据 MV 方法，此时最终得到的标记是错误的[16,17]。此外，MV 方法只根据当前样例的标记进行投票，它忽略了邻居样例的信息。现实中，众包数据还会出现有的样例比较难以识别，导致没有标记员标记，或者某个样例的标记全部缺失，此时，根据 MV 方法的原则是无法返回一个明确的答案的。

2.2　W-Knv 方法

针对以上问题，本文提出了加权 K 近邻方法(Weighted k nearest neighbor voting，缩写为 W-Knv)，如式（2）所示。

$$v_k(x) = \arg\max_{c\in\Omega}\ v_k(c\,|\,x) \tag{2}$$

其中：$v_k(c\,|\,x) = \frac{1}{|S_x|+\bar{\alpha}}[|S_x|v(c\,|\,x)+\alpha_c^x]$，$\alpha_c^x = \frac{1}{k}\sum_{i=1}^{k}\alpha_i v(c\,|\,x_i)$，$x_i \in N_{K(x)}$。

本文用向量 $\boldsymbol{\alpha}$ 表示样例 X 邻居的标记的权重。由于越近，说明样例的关系越紧密，所以初始化 $\boldsymbol{\alpha}=[k,k-1,k-2,\cdots\cdots,1]$，$\bar{\boldsymbol{\alpha}}$ 表示向量 $\boldsymbol{\alpha}$ 内元素的均值。仔细观察上式，随着 $|S_x|$ 的增大，$\bar{\boldsymbol{\alpha}}$ 对 W-Knv 算法的影响就越小。当 $\boldsymbol{\alpha}$ 内元素的值全部为 0，也就表示样例的近邻对结果基本没有影响。此时，本文提出的 W-Knv 方法就与 MV 方法完全一致。随着 $\bar{\boldsymbol{\alpha}}$ 值的增大，邻居标记的重要性也随着增大。同时，若样例中出现标记缺失或标记不平衡时，本文提出的 W-Knv 方法可以根据样例邻居的标记科学地返回一个明确的答案。

本文提出的 W-Knv 方法主要有以下优点：

a)由于现实生活中的众包数据是有标记缺失、噪声等影响的。K 近邻投票分类方法通过考虑样例邻居之间的标记，

可以减轻一部分这方面的影响，无论标记是否缺失，或无标记，都可以给出一个明确的答案。

b)传统的多数投票分类方法比较适合样例标记很均衡的情况，但实际中标记员的知识水平参差不齐，所以就会造成样例的标记个数不均衡，从而使传统的投票方法获取的样例标记不够多，产生准确率比较低的答案。而 K 近邻投票分类方法很容易获取足够多的标记。

c)在实际中，样例之间通常是有一定关系的，属于同一类的样例很可能是聚集在一起。多数投票分类方法只是根据样例的标记来给出答案，忽略了样例之间的关系。K 近邻综合了样例邻居的标记信息，从而可以给出更准确的答案。

d)通过调节向量 α 的值，可以控制样例邻居标记的权重，这样可以排除一些错误标记的干扰，从而使本文提出的 K 近邻投票分类方法的性能更好。

3 实验结果与分析

3.1 实验数据集和参数设置

在本文实验中采用 CCUDS、CNAE、Drift、Ecoli、Yale、Chess、Movements、Soybean 这八个数据集来验证提出的 W-Knv 算法。这些数据集均来自 UCI 数据集[18]。数据集的详细信息如表 2 所示。

表 2 数据集详细信息

Table 2 Data set details

数据集	样本数	属性数	类数
CCUDS	1994	101	10
CNAE	1080	856	9
Drift	1244	129	6
Ecoli	336	343	8
Yale	165	1024	15
Chess	3196	36	2
Movements	360	90	15
Soybean	307	35	19

为了从各个方面验证本文提出的 W-KNV 算法的性能，设置如下参数：

a) 平均标记数 $|\bar{S}_x|$。它的大小影响着总标记数。由于实际中每个样例的标记数并不都是一样的（可能出现无标记或标记缺失的情况），设置它为 3 或 5，以此来模拟实际的标记数。

b) 贝塔分布的参数 con。本文通过 $|S|=|2\times|\bar{S}_x|\times\lambda|$ 来得到每个样例的标记数，这个式子决定了构建每个样例的标记数。其中：$|S|$ 代表某个样例的标记数；$\lambda\sim B(con,con)$，这样每个样例的标记数就可以取到 $(0,\cdots\cdots,2|\bar{S}_x|)$，从而使得每个样例的标记数不固定，可以更加真实地模拟真实众包数据。

c) 可靠性参数 rel。这个参数表示标记员标记正确一个样例的概率。由于实际中标记员标错的概率并不是很大，所以本文设置它的值范围为 $rel\in(0.5,\cdots\cdots1)$。

接下来，首先要把这些数据集进行处理来生成本文所需要的标记，即众包数据[19]。人工生成众包数据的具体过程如下：

a) 本文首先通过随机森林算法来对数据进行分类，然后利用分类后得到的预测标签和数据本身的真实标签来生成一个混淆矩阵 $\boldsymbol{M}$。$\boldsymbol{M}$ 矩阵如下所示：

$$\boldsymbol{M}=\begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,c} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,c} \\ \cdots & \cdots & \cdots & \cdots \\ m_{c,1} & m_{c,2} & \cdots & m_{c,c} \end{bmatrix}$$

其中：$\boldsymbol{M}\in R^{c\times c}$，c 代表数据的类别数。本文用 $M_{i,j}$ 表示矩阵 $\boldsymbol{M}$ 中的元素。$M_{i,j}=pr(\text{标记员给的标记是}j\,|\,\text{该样例的真实标签是}i)$。

b) $\boldsymbol{M}$ 矩阵构建完成后，再根据如下规则构建 $\boldsymbol{R}$ 矩阵，$\boldsymbol{R}$ 矩阵的大小和 $\boldsymbol{M}$ 矩阵是相同的：(a)令 $\boldsymbol{R}$ 矩阵的主对角元素 $R_{c,c}=rel$；(b)设置 $\boldsymbol{R}$ 矩阵的其他元素值为

$$R_{c,c'}=\begin{cases}\dfrac{1-rel}{c-1} & \text{if } \boldsymbol{M}_c=\boldsymbol{0} \text{ except for } M_{c,c} \\ \dfrac{M_{c,c'}(1-rel)}{\sum_{c'\neq c}M_{c,c'}} & otherwise\end{cases} \tag{3}$$

其中：$c\neq c'$；$\boldsymbol{M}_c$ 代表 $\boldsymbol{M}$ 矩阵的第 c 行；$\boldsymbol{0}$ 代表全为零；$M_{c,c}$ 代表对角线元素。

$\boldsymbol{R}$ 矩阵构建好了之后，开始构建众包数据的标记。当给定一个样例之后，若该样例属于第 c 类，就从 $\boldsymbol{R}$ 矩阵的第 c 行中抽取 $|S|$ 次，把抽取的元素作为该样例的标记集合。这样对每个样例依次进行，就生成众包数据了。

以上的种种做法，都使得本文生成的众包数据更加贴近于现实生活中的真实数据。

3.2 实验结果与分析

接下来本文从各个方面来测试本文算法的性能。本章所有实验都是在 Windows 7 环境下的 MATLAB 2014a 平台上进行的。

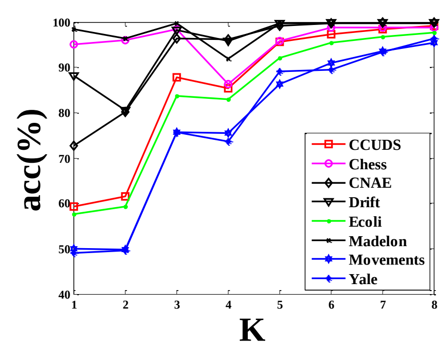


图 1 W-Knv 取不同 K 值的准确率

Fig. 1 W-Knv takes the accuracy of different K values

如图 1 所示，在参数 rel=0.6，con=1，$\alpha=[5,4,3,2,1]$，$\bar{\alpha}=3$，$|\bar{S}_x|=5$ 的设置下，W-Knv 方法在 K 取不同值的情况下的准确率。可以看出，当 $k\geq5$ 时，W-Knv 方法的准确率变化不大。因此，在之后的实验中设定 K 值为 5。为了测试参数 con 对算法 MV 和 W-Knv 的影响，固定其他参数值为：K=5，rel=0.6，$\alpha=[5,4,3,2,1]$，$\bar{\alpha}=3$，$|\bar{S}_x|=3$ or 5。

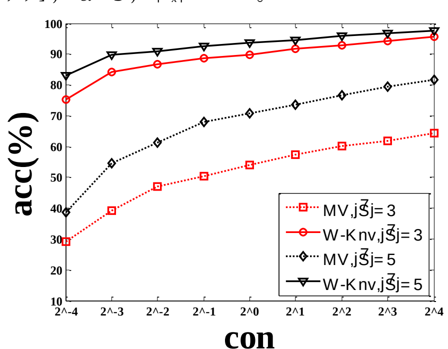


图 2 数据集 OCCUDS

Fig. 2 Data set OCCUDS

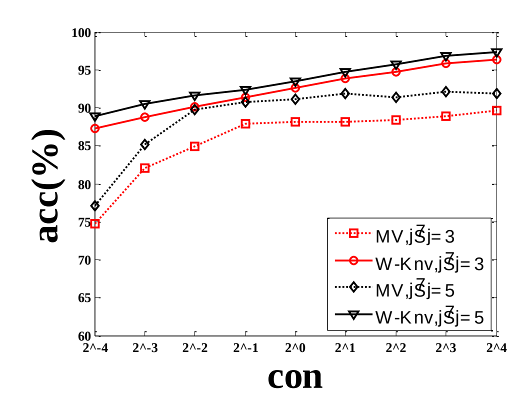


图 3 数据集 Chess

Fig. 3 Data set Chess

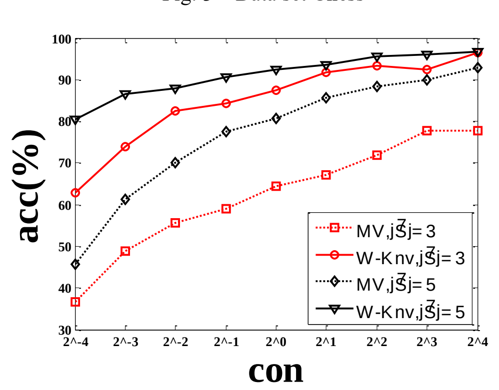


图 4 数据集 Ecoli

Fig. 4 Data set Ecoli

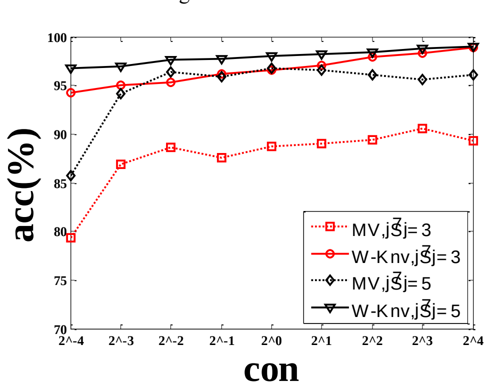


图 6 数据集 Madelon

Fig. 6 Data set Madelon

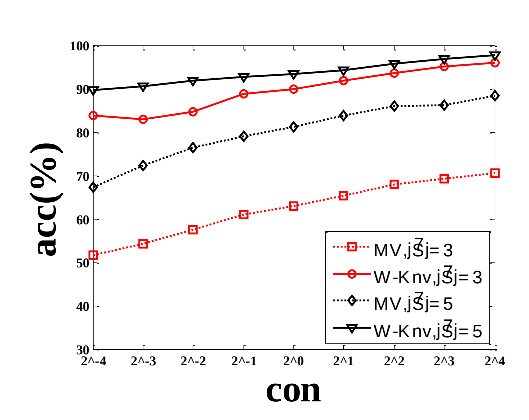


图 8 数据集 CNAE

Fig. 8 Data set CNAE

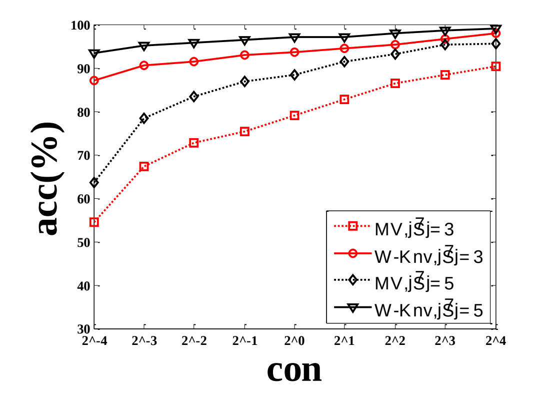


图 5 数据集 Drift

Fig. 5 Data set Drift

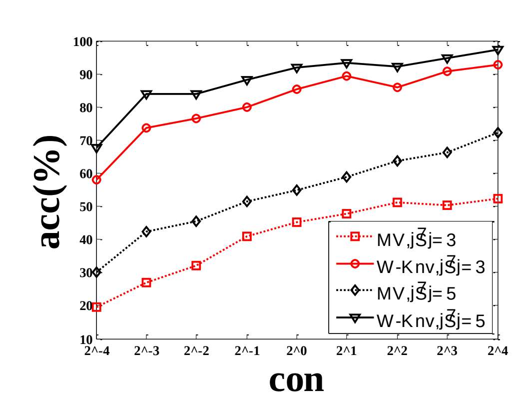


图 7 数据集 Movements

Fig. 7 Dataset Movements

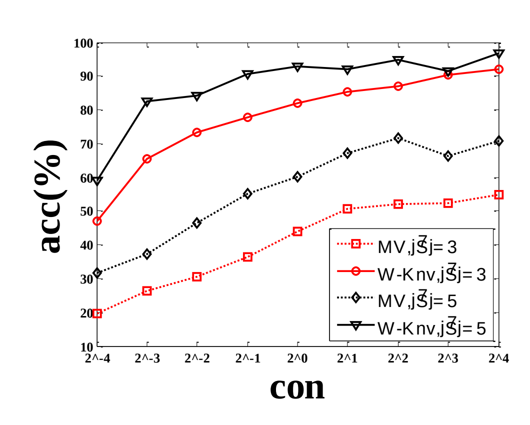


图 9 数据集 Yale

Fig. 9 Data set Yale

由于本章设置贝塔分布的两个参数是相同的值，所以当参数 con 的值取很小的时候，每个样例的标记数就会趋向于零。而当参数的值很大的时候，每个样例的标记数就会趋向于 $2|S_x|$。图 2~9 展示了 MV 算法和 W-Knv 算法在八个数据集上的准确率。可以发现，随着参数 con 值的增加，准确率都是递增的；当 con 值较小时，标记数比较少，此时 MV 算法通常不能给出较为准确的答案，受影响比较大。同时本章设置 $|\tilde{S}_x| = 3$ or 5，这样也间接调节了标记的个数。通过这八个图可以看出，随着标记数的变化，W-Knv 算法的性能都比 MV 算法要好。此外，通过改变参数 rel 的值，本文也做了一些实验。

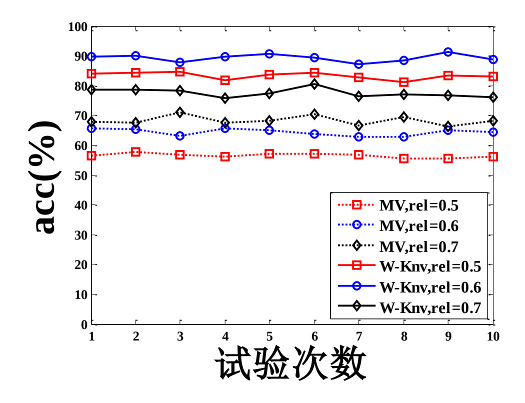


图 10　数据集 OCCUDS

Fig. 10　Data set OCCUDS

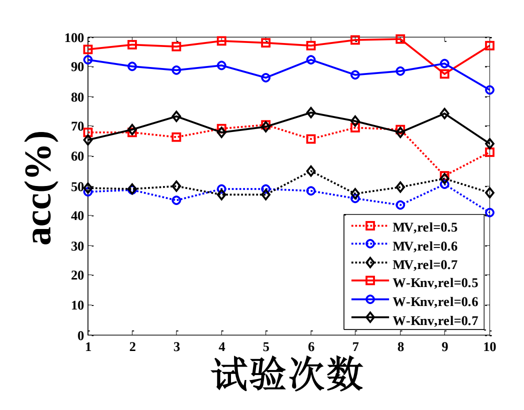


图 11　数据集 Ecoli

Fig. 11　Data set Ecoli

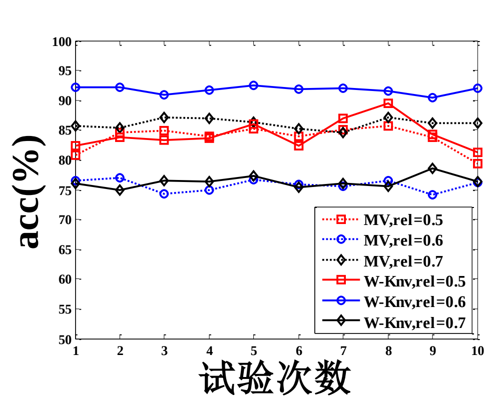


图 12　数据集 Chess

Fig. 12　Data set Chess

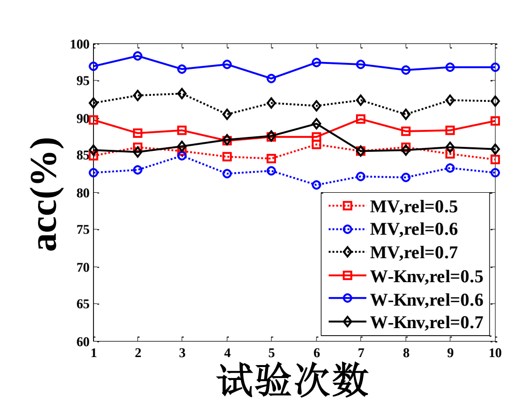


图 13　数据集 Madelon

Fig. 13　Data set Madelon

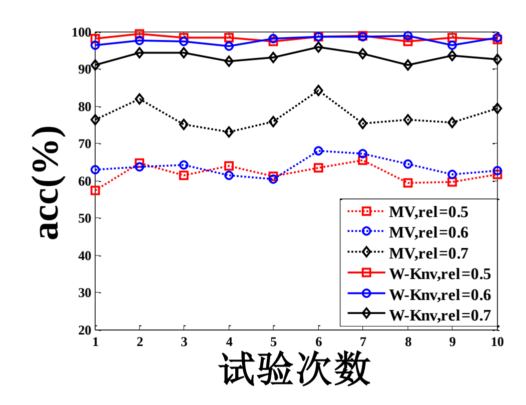


图 14　数据集 CNAE

Fig. 14　Data set CNAE

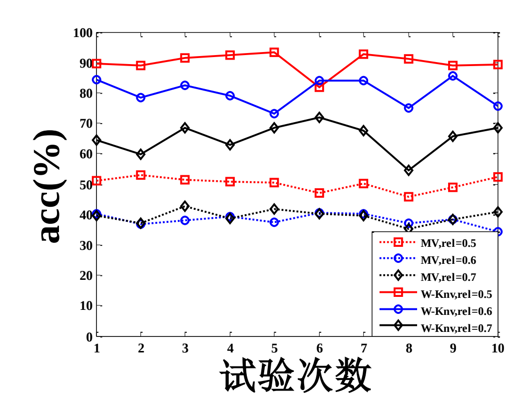


图 15　数据集 Movements

Fig. 15　Dataset movements

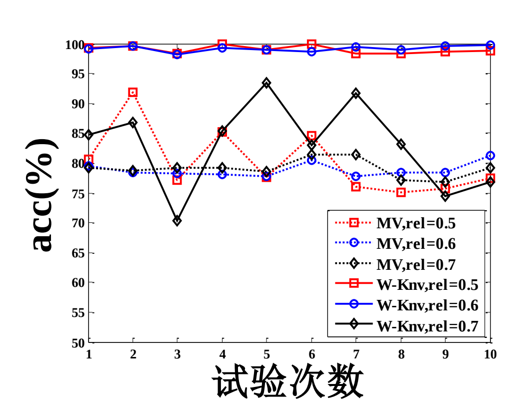


图 16　数据集 Drift

Fig. 16　Dataset drift

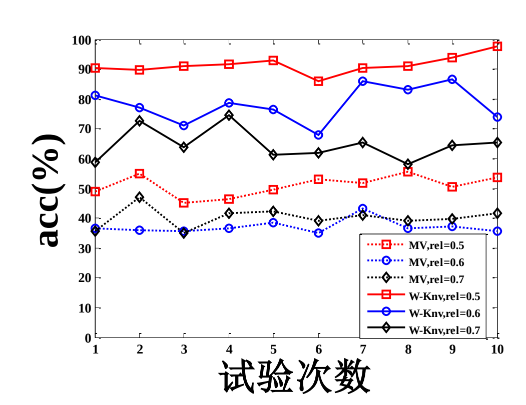


图 17　数据集 Yale

Fig. 17　Data set Yale

如图 10~17 所示，在参数 K=5，con=1，$\alpha=[5,4,3,2,1]$，$\bar{\alpha}=3$，$|\bar{S}_x|=5$ 的设置下，本文在八个数据集上分别进行了 10 次实验，来验证不同的 rel 值对算法的影响。

通过对实验结果的分析，发现准确率并没有随着 rel 值的增大而提高。但大多数情况下，本文提出的 W-Knv 算法准确率都比 MV 算法要好。通过调节参数 rel 的值来调整标记员的水平，以此来提升算法的性能。

从(1)式和(2)中可以看出，无论是传统的 MV 算法，还是本文提出的 W-Knv 算法，时间复杂度都主要与 $|S|$ 有关。由于都是线性的计算，所以它们两种算法的时间复杂度都是 O(n)。在实际情况中，本文提出的 W-Knv 算法需要计算近邻，所以计算次数会多一点。但只是多了一些线性的计算，总体的算法时间复杂度还是 O(n)。

当本文把式（2）中向量 α 里的元素的值全设置为 k 时，此时表示所有近邻的权重是一样的，即把传统的 K 近邻算法思想运用到众包数据中。很明显，这样的效果没有本文的加权 K 近邻更符合常理。在性能上，本文的加权 K 近邻可以通过调节向量 α 的值来调节权重，传统的 K 近邻不可以。可以说，本文在众包数据上的加权 K 近邻标签聚合算法是传统 K 近邻算法思想的一种改进运用。

综上所述，本文提出的 W-Knv 算法之所以取得了较好的性能，主要有以下三个原因：a)在部分样例标记缺失或较少的时候，也能返回一个较好的答案，适用性更强；b)不仅考虑了样例的标记，还考虑了样例之间的关系，以此来得到更加精确的答案；c)通过给样例的近邻来设置权重，可以有效地去除不准确的标记，使算法性能更好。

4 结束语

本文通过考虑样例的 K 近邻标记，提出了一种在众包学习中的加权 K 近邻投票分类算法。即通过 K 近邻来找出样例的邻居，根据距离由近到远对邻居样例所对应的标记设置权重，越近的权重越大。以此来使算法的性能更好。该算法保留了多数投票的简单性，在一定程度上可以代替多数投票分类法。经实验结果证实，本文算法取得了很好的效果。在今后的工作中，本文尝试通过不同的分类算法和概率估计来进行改善。

参考文献：

[1] Deng Zhenyun, Zhu Xiaofeng, Cheng Debo, *et al.* Efficient KNN classification algorithm for big data [J]. Neurocomputing, 2016, 195 (C): 143-148.

[2] Zhang Shichao, Li Xuelong, Zong Ming, *et al.* Efficient KNN classification with different numbers of nearest neighbors [J]. IEEE Trans on Neural Networks & Learning Systems, 2018, 29 (5): 1774-1785.

[3] Zhu Xiaofeng, Zhang Shichao, Jin Zhi, *et al.* Missing value estimation for mixed-attribute datasets [J]. IEEE Trans on Knowledge and Data Engineering, 2011, 23 (1): 110-121.

[4] Deng Jia, Dong Wei, Socher R, *et al.* ImageNet: A large-scale hierarchical image database [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S. l.] : IEEE Press, 2009: 248-255.

[5] Zhu Xiaofeng, Li Xuelong, Zhang Shichao. Block-row sparse multi-view multi-label learning for image classification [J]. IEEE Trans on Cybernetics, 2016, 46 (2): 450-461.

[6] Qin Yongsong, Zhang Shichao, Zhu Xiaofeng, *et al.* Semi-parametric optimization for missing data imputation [J]. Applied Intelligence, 2007, 27 (1): 79-88.

[7] Zhang Shichao, Zong Ming, Sun Ke, *et al.* Efficient KNN algorithm based on graph sparse reconstruction [M]// Advanced Data Mining and Applications. Germany: Springer Press, 2014: 356-369.

[8] 郝建柏，陈贤富，黄双福，等. 一种基于模糊近邻标签传递的半监督分类算法 [J]. 微电子学与计算机, 2010, 27 (2): 30-33. (Hao Jianbai, Chen Xianfu, Huang Shuangfu, *et al.* A semi-supervised classification algorithm based on fuzzy near-neighbor label transfer [J]. Microelectronics & Computer, 2010, 27 (2): 30-33.)

[9] Zhang Minling, Zhou Zhihua. ML-KNN: a lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007, 40 (7): 2038-2048.

[10] Zhang Shichao, Li Xuelong, Zong Ming, *et al.* Learning k, for KNN classification [J]. ACM Trans on Intelligent Systems & Technology, 2017, 8 (3): 43.

[11] Zhang, Shichao. KNN-CF approach: incorporating certainty factor to kNN classification [J]. IEEE Intelligent Informatics Bulletin 2010 11 (1) , 25–34

[12] 邓生雄，雒江涛，刘勇，等. 集成随机森林的分类模型 [J]. 计算机应用研究, 2015, 32 (6): 1621-1624. (Deng Shengxiong, Yan Jiangtao, Liu Yong, *et al.* Classification model of integrated random forests [J]. Application Research of Computers, 2015, 32 (6): 1621-1624.)

[13] 韩慧，毛锋，王文渊. 数据挖掘中决策树算法的最新进展 [J]. 计算机应用研究, 2004, 21 (12): 5-8. (Han Hui, Mao Feng, Wang Wenyuan. Recent advances in decision tree algorithms in data mining [J]. Application Research of Computers, 2004, 21 (12): 5-8.

[14] Snow R, O'Connor B, Jurafsky D, *et al.* Cheap and fast-but is it good?: evaluating non-expert annotations for natural language tasks [C]// Proc of Conference on Empirical Methods in Natural Language Processing. [S.l.]:Association for Computational Linguistics, 2008: 254-263.

[15] Hu Huiqi, Zheng Yudian, Bao Zhifeng, *et al.* Crowdsourced POI labelling: location-aware result inference and task assignment [C]// Proc of IEEE International Conference on Data Engineering. [S. l.] : IEEE Press, 2016: 61-72.

[16] Ouyang R W, Kaplan L, Martin P, *et al.* Debiasing crowdsourced quantitative characteristics in local businesses and services [C]// Proc of the 14th International Conference on Information Processing in Sensor Networks. [S. l.] : ACM Press, 2015: 190-201.

[17] Li Guoliang, Wang Jiannan, Zheng Yudian, *et al.* Crowdsourced data management: a survey [J]. IEEE Trans on Knowledge & Data Engineering, 2016, 28 (9): 2296-2319.

[18] UCI repository of machine learning datasets [EB/OL]. [2016-05-27] http://archive. ics. uci. edu/ml/.

[19] Cao C C, She Jieying, Tong Yongxin, *et al.* Whom to ask?: Jury selection for decision making tasks on micro-blog services [J]. Proceedings of the Vldb Endowment, 2012, 5 (11): 1495-1506.